VECTOR UNIFORM CRAMER-RAO LOWER BOUND

Yonina C. Eldar

Department of Electrical Engineering
Technion–Israel Institute of Technology, Haifa, Israel
Email: yonina@ee.technion.ac.il.

ABSTRACT

We develop a uniform Cramer-Rao lower bound (UCRLB) on the total variance of any estimator of an unknown deterministic vector of parameters, with bias gradient matrix whose norm is bounded by a constant. We consider two different measures of norm, leading to two corresponding bounds. When the observations are related to the unknown vector through a linear Gaussian model, Tikhonov regularization and the shrunken estimator are shown to achieve the UCRLB. For more general models, we show that the penalized maximum likelihood estimator with a suitable penalizing function *asymptotically* achieves the UCRLB.

1. INTRODUCTION

A common approach to developing well behaved estimators in overparameterized estimation problems is to use regularization techniques, with generally measure both the fit to the observed data and the physical plausibility of the estimate. Often, the use of regularization can reduce the variance of the estimator at the expense of increasing the bias, so that the design of such estimators is typically subject to a tradeoff between variance and bias.

We consider the class of estimation problems in which we seek to estimate an unknown deterministic vector $\mathbf{x}_0$ from measurements $\mathbf{y}$, where the relationship between $\mathbf{y}$ and $\mathbf{x}_0$ is described by the joint pdf $p(\mathbf{y};\mathbf{x}_0)$ of $\mathbf{y}$ characterized by $\mathbf{x}_0$.

The total variance of any unbiased estimator of $\mathbf{x}_0$ is bounded by the Cramer-Rao lower bound (CRLB) [1]. If $\mathbf{y}$ is related to $\mathbf{x}_0$ through a linear Gaussian model, then the maximum likelihood (ML) estimate of $\mathbf{x}_0$ achieves the CRLB. Furthermore, when $\mathbf{x}_0$ is estimated from independent identically distributed (iid) measurements, under suitable regularity assumptions on $p(\mathbf{y};\mathbf{x}_0)$, the ML estimator is asymptotically unbiased and achieves the CRLB [1].

Since estimators resulting from regularization methods are typically biased, their variance cannot be bounded by the CRLB. The total variance of any estimator with a *given* bias gradient is bounded by the *biased CRLB* [2]. However, in applications it may not be obvious how to choose a particular bias gradient. Therefore, it would be useful to have a lower bound on the smallest attainable variance using any estimator whose bias gradient belongs to a suitable class. A bound of this form was first developed by Hero *et al.* [3] for estimating a *scalar* function of a deterministic vector parameter. They propose the *uniform CRLB (UCRLB)*, which is a bound on the smallest variance that can be achieved using any estimator with bias gradient whose norm is bounded by a constant.

In this paper we extend the results of [3] in two ways. First, in Section 3, we derive a UCRLB for vector parameters. Specifically, we derive bounds on the total variance of any estimator $\hat{\mathbf{x}}$ of $\mathbf{x}_0$, with bias gradient matrix whose norm is bounded. We consider two different matrix norms which lead to two lower bounds: the Frobenius norm corresponding to an average bias gradient measure, and the spectral norm corresponding to a worst case bias gradient measure.

Second, we develop estimators that achieve the UCRLB. Specifically, in Section 4 we consider a linear Gaussian model, and derive linear estimators of $\mathbf{x}_0$ that achieve the UCRLB. In Section 5, we consider the case of estimating $\mathbf{x}_0$ from iid vector measurements, and develop a class of *penalized maximum likelihood (PML) estimators* that asymptotically achieve the UCRLB.

Proofs of theorems, which are omitted here for brevity, can be found in [4].

2. BIASED CRAMER-RAO LOWER BOUND

We denote vectors and matrices by boldface lowercase and uppercase letters, respectively. The Hermitian conjugate is denoted by $(\cdot)^*$. The true value of an unknown vector (scalar) $\mathbf{x}$ (x) is denoted by $\mathbf{x}_0$ (x_0). $\partial f(\mathbf{x}_0)/\partial \mathbf{x}$ denotes the gradient of the function $f(\mathbf{x})$ evaluated at the point $\mathbf{x}_0$, and is a row vector with j element equal to $\partial f(\mathbf{x}_0)/\partial x_j$. The gradient of a vector $\partial \mathbf{b}(\mathbf{x}_0)/\partial \mathbf{x}$ is a matrix, with ijth element equal to $\partial b_i(\mathbf{x}_0)/\partial x_j$. The largest eigenvalue of a matrix $\mathbf{A}$ is denoted by $\lambda_{\max}(\mathbf{A})$, and $\mathcal{N}(\mathbf{m},\mathbf{C})$ denotes the Gaussian distribution with mean $\mathbf{m}$ and covariance matrix $\mathbf{C}$.

Let $\hat{\mathbf{x}}$ denote an arbitrary estimator of $\mathbf{x}_0 \in \mathbb{C}^m$ from the observations $\mathbf{y} \in \mathbb{C}^n$, with bias $\mathbf{b}(\mathbf{x}_0) = E(\hat{\mathbf{x}}) - \mathbf{x}_0$, and covariance $\mathbf{C}_{\hat{\mathbf{x}}} = E\{[\hat{\mathbf{x}} - E(\hat{\mathbf{x}})][\hat{\mathbf{x}} - E(\hat{\mathbf{x}})]^*\}$. Under suitable regularity conditions on $p(\mathbf{y};\mathbf{x})$ [1], $\mathbf{C}_{\hat{\mathbf{x}}}$ must satisfy

$$\mathbf{C}_{\hat{\mathbf{x}}} \geq (\mathbf{I}+\mathbf{D})\,\mathbf{J}^{-1}\,(\mathbf{I}+\mathbf{D})^* \stackrel{\triangle}{=} \mathbf{C}(\mathbf{D}), \tag{1}$$

where $\mathbf{J}$ is the Fisher information matrix defined by

$$\mathbf{J} = E\left\{\left[\frac{\partial \log p(\mathbf{y};\mathbf{x}_0)}{\partial \mathbf{x}}\right]^*\left[\frac{\partial \log p(\mathbf{y};\mathbf{x}_0)}{\partial \mathbf{x}}\right]\right\}, \tag{2}$$

and is assumed to be nonsingular, and $\mathbf{D}$ is the bias gradient matrix defined by $\mathbf{D} = \partial \mathbf{b}(\mathbf{x}_0)/\partial \mathbf{x}$.

For a given bias gradient $\mathbf{D}$, the total variance that is achievable using any linear or nonlinear estimator with this bias gradient is bounded below by $\mathrm{Tr}(\mathbf{C}(\mathbf{D}))$, where the total variance $\sum_{i=1}^m E\left\{[\hat{x}_i - E(\hat{x}_i)]^2\right\}$ is the sum of the variances in estimating the individual components of $\mathbf{x}_0$. Typically, in estimation problems, there are two conflicting objectives that we would like to minimize: We would like to choose an estimator $\hat{\mathbf{x}}$ to achieve the smallest possible total variance *and* the smallest possible bias. However, generally, minimizing the bias results in an increase in variance and *vice versa*. To quantify the best achievable performance of any estimator $\hat{\mathbf{x}}$

of $\mathbf{x}_0$ taking both the bias and the total variance into account, we choose to minimize the total variance

$$C(\mathbf{D}) = \mathrm{Tr}(\mathbf{C}(\mathbf{D})) = \mathrm{Tr}\left((\mathbf{I}+\mathbf{D})\mathbf{J}^{-1}(\mathbf{I}+\mathbf{D})^*\right), \quad (3)$$

subject to a constraint on the bias gradient matrix $\mathbf{D}$.

In our development we consider two measures of bias gradient: an average bias gradient measure corresponding to a weighted squared Frobenius norm,

$$D_{\text{AVG}} = \mathrm{Tr}(\mathbf{D}^*\mathbf{D}\mathbf{W}), \quad (4)$$

where $\mathbf{W}$ is an arbitrary nonnegative definite weighting matrix, and a worst case bias gradient measure corresponding to a weighted squared spectral norm,

$$D_{\text{WC}} = \max_{\mathbf{z}\in\mathbb{C}^m, \|\mathbf{z}\|=1} \mathbf{z}^*\mathbf{S}\mathbf{D}^*\mathbf{D}\mathbf{S}\mathbf{z}, \quad (5)$$

for some nonnegative definite matrix $\mathbf{S}$.

3. VECTOR UCRLB

We first consider the problem of minimizing (3) subject to $D_{\text{AVG}} \le \gamma$. If $\gamma \ge \mathrm{Tr}(\mathbf{W})$, then we can choose $\mathbf{D} = -\mathbf{I}$ which results in $C(\mathbf{D}) = 0$. The average bound for the case $\gamma < \mathrm{Tr}(\mathbf{W})$ is given in the following theorem.

Theorem 1 *Let $\mathbf{x}_0$ denote an unknown deterministic vector, let $\mathbf{y}$ denote measurements of $\mathbf{x}_0$, and let $p(\mathbf{y};\mathbf{x}_0)$ denote the pdf of $\mathbf{y}$ characterized by $\mathbf{x}_0$. Let $\mathbf{W}$ be a nonnegative Hermitian weighting matrix. Then the total variance $C = C(\mathbf{D})$ defined by (3) of any estimator of $\mathbf{x}_0$ with bias gradient matrix $\mathbf{D}$ such that $Tr(\mathbf{D}^*\mathbf{D}\mathbf{W}) \le \gamma < Tr(\mathbf{W})$ satisfies*

$$C \ge \alpha^2 Tr\left((\mathbf{I}+\alpha\mathbf{W}\mathbf{J})^{-1}\mathbf{W}\mathbf{J}\mathbf{W}(\mathbf{I}+\alpha\mathbf{J}\mathbf{W})^{-1}\right),$$

where $\alpha > 0$ is the unique scalar for which

$$Tr\left((\mathbf{I}+\alpha\mathbf{W}\mathbf{J})^{-1}\mathbf{W}(\mathbf{I}+\alpha\mathbf{J}\mathbf{W})^{-1}\right) = \gamma.$$

We next consider the problem of minimizing (3) subject to $D_{\text{WC}} \le \gamma$. If $\gamma \ge \lambda^2_{\max}(\mathbf{S})$, then we can choose $\mathbf{D} = -\mathbf{I}$ which results in $C(\mathbf{D}) = 0$. In the general case, the worst-case bound can be found as a solution to a semidefinite programming problem [5], which is a convex optimization problem that can be solved very efficiently.

Theorem 2 *Let $\mathbf{x}_0$ denote an unknown deterministic vector, let $\mathbf{y}$ denote measurements of $\mathbf{x}_0$, and let $p(\mathbf{y};\mathbf{x}_0)$ denote the pdf of $\mathbf{y}$ characterized by $\mathbf{x}_0$. Let $\mathbf{S}$ be an arbitrary nonnegative definite matrix. Then the total variance $C = C(\mathbf{D})$ of any estimator of $\mathbf{x}_0$ with bias gradient matrix $\mathbf{D}$ such that $\max_{\mathbf{z}\in\mathbb{C}^m,\|\mathbf{z}\|=1} \mathbf{z}^*\mathbf{S}\mathbf{D}^*\mathbf{D}\mathbf{S}\mathbf{z} \le \gamma < \lambda^2_{\max}(\mathbf{S})$ satisfies $C \ge C_{\min}$ where C_{min} is the solution to the semidefinite program*

$$C_{\min} = \min_{t,\mathbf{D}} t$$

subject to

$$\begin{bmatrix} t & \mathbf{g}^* \\ \mathbf{g} & \mathbf{I} \end{bmatrix} \succeq 0,$$

$$\begin{bmatrix} \gamma\mathbf{I} & \mathbf{S}\mathbf{D}^* \\ \mathbf{D}\mathbf{S} & \mathbf{I} \end{bmatrix} \succeq 0,$$

with $\mathbf{g} = vec(\mathbf{J}^{-1/2}(\mathbf{I}+\mathbf{D})^*)$.

If $\mathbf{S} = \sum_{i=1}^m \beta_i \mathbf{q}_i\mathbf{q}_i^*$ for some $\beta_i > 0$, where $\mathbf{q}_i$ are the eigenvectors of $\mathbf{J}$, then

$$C_{\min} = Tr\left(\left(\mathbf{I}-\sqrt{\gamma}\mathbf{S}^{-1}\right)^2\mathbf{P}\mathbf{J}^{-1}\right).$$

Here $\mathbf{P} = \sum_{i:\beta_i^2>\gamma} \mathbf{q}_i\mathbf{q}_i^*$ is the orthogonal projection onto the space spanned by the eigenvectors of $\mathbf{S}$ corresponding to eigenvalues $\beta_i^2 > \gamma$. If, in addition, $\mathbf{S} = \mathbf{I}$, then

$$C_{\min} = Tr\left((1-\sqrt{\gamma})^2\mathbf{J}^{-1}\right).$$

Theorems 1 and 2 characterize the smallest possible total variance of any estimator with bias gradient matrix whose norm is bounded by a constant. However, the theorems do not guarantee that there exists estimators achieving these bounds. In the next section we show that in the case of a linear Gaussian model, both bounds are achievable using a linear estimator. In Section 5, we consider more general, not necessarily Gaussian models, and develop a class of estimators that *asymptotically* achieve the UCRLB.

4. LINEAR GAUSSIAN MODEL

Consider the estimation problems represented by the model

$$\mathbf{y} = \mathbf{H}\mathbf{x}_0 + \mathbf{n}, \quad (6)$$

where $\mathbf{x}_0 \in \mathbb{C}^m$ is an unknown deterministic vector, $\mathbf{H}$ is a known $n \times m$ matrix with rank m, and $\mathbf{n} \in \mathbb{C}^n$ is a zero-mean Gaussian random vector with positive definite covariance $\mathbf{C}_n$. For this model, the Fisher information matrix is $\mathbf{J} = \mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{H} \triangleq \mathbf{Q}$.

It is straightforward to show that the estimator

$$\hat{\mathbf{x}} = \begin{cases} (\mathbf{W}\mathbf{Q}+\delta\mathbf{I})^{-1}\mathbf{W}\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{y}, & \gamma < \mathrm{Tr}(\mathbf{W}); \\ 0, & \gamma \ge \mathrm{Tr}(\mathbf{W}), \end{cases} \quad (7)$$

where the regularization parameter $\delta > 0$ is chosen such that $\mathrm{Tr}\left((\mathbf{I}+(1/\delta)\mathbf{W}\mathbf{Q})^{-1}\mathbf{W}(\mathbf{I}+(1/\delta)\mathbf{Q}\mathbf{W})^{-1}\right) = \gamma$, achieves the average UCRLB of Theorem 1, so that among all estimators with bias gradient $\mathbf{D}$ satisfying $\mathrm{Tr}(\mathbf{D}^*\mathbf{D}\mathbf{W}) \le \gamma$, this estimator results in the smallest possible total variance. Note, that the estimator $\hat{\mathbf{x}}$ of (7) is equal to the Tikhonov regularizer [6], which is widely used for solving inverse problems and ill-conditioned least-squares problems.

Similarly, among all estimators with bias gradient $\mathbf{D}$ satisfying $\mathbf{z}^*\mathbf{S}\mathbf{D}^*\mathbf{D}\mathbf{S}\mathbf{z} \le \gamma < \lambda^2_{\max}(\mathbf{S})$ for all $\mathbf{z} \in \mathbb{C}^m$ such that $\mathbf{z}^*\mathbf{z} = 1$, where $\mathbf{S}$ is a positive definite matrix that commutes with $\mathbf{Q}$ and has eigenvalues β_i, the estimator that results in the smallest possible total variance is

$$\hat{\mathbf{x}} = \begin{cases} (\mathbf{I}-\sqrt{\gamma}\mathbf{S}^{-1})\mathbf{P}\mathbf{Q}^{-1}\mathbf{H}^*\mathbf{C}_n^{-1}\mathbf{y}, & \gamma < \lambda^2_{\max}; \\ 0, & \gamma \ge \lambda^2_{\max}, \end{cases} \quad (8)$$

where $\mathbf{P}$ is an orthogonal projection onto the space spanned by the eigenvectors of $\mathbf{S}$ corresponding to eigenvalues $\beta_i^2 > \gamma$. The estimator $\hat{\mathbf{x}}$ of (8) with $\mathbf{S} = \mathbf{I}$ is equal to the shrunken estimator proposed by Mayer and Willke [7], which is simply a scaled version of the least-squares estimator. For more general choices of $\mathbf{S}$, the estimator of (8) can be viewed as a generalization of the shrunken estimator.

5. ASYMPTOTIC OPTIMALITY OF PML ESTIMATION

In general, there is no guarantee that an estimator achieving the UCRLB exists. We have seen that for the linear Gaussian model, the average UCRLB is achieved by Tikhonov regularization, which also maximizes the penalized log-likelihood function $p(\mathbf{y};\mathbf{x}) - \beta \mathbf{x}^* \mathbf{W} \mathbf{x}$, where $p(\mathbf{y};\mathbf{x}) \sim \mathcal{N}(\mathbf{H}\mathbf{x}, \mathbf{C}_n)$. A similar result holds for the shrunken estimator.

We now demonstrate that this optimality property of the PML estimator is more general. Specifically, we show that the PML estimator *asymptotically* achieves the UCRLB for many other statistical models. To this end, we first develop the asymptotic bias and variance of the PML estimator for a general class of penalizing functions. We then show that in many cases we can choose the penalizing function such that the PML estimator asymptotically achieves the UCRLB.

5.1 Asymptotic Properties of the PML Estimator

The PML estimate of $\mathbf{x}_0$, denoted $\hat{\mathbf{x}}^{\text{PML}}$, is chosen to maximize the penalized log-likelihood $\log p(\mathbf{y};\mathbf{x}) - \beta R(\mathbf{x})$, where $\beta > 0$ is a regularization parameter, and $R(\mathbf{x})$ is a penalizing function. Although many different choices of $R(\mathbf{x})$ have been proposed in the literature [8, 9], no general assertions of optimality are known for these different choices.

In the case in which we estimate $\mathbf{x}_0$ from N iid (vector) measurements $\mathbf{y}_1, \ldots, \mathbf{y}_N$, $\hat{\mathbf{x}}^{\text{PML}}$ is chosen to maximize

$$PL(x) = \sum_{i=1}^{N} \log p(\mathbf{y}_i;\mathbf{x}) - \beta_N R(\mathbf{x}), \tag{9}$$

where β_N is a regularization parameter that may depend on N. In our derivations, we assume that $\beta_N/N \to \beta_0$ for some constant β_0 as $N \to \infty$. Under suitable regularity conditions, we have the following theorem:

Theorem 3 *Let $\mathbf{x}_0$ denote an unknown deterministic vector, let $\mathbf{y}_1, \ldots, \mathbf{y}_N$ denote N iid measurements of $\mathbf{x}_0$, and let $\hat{\mathbf{x}}^{\text{PML}}$ denote the PML estimator of $\mathbf{x}_0$ from $\mathbf{y}_1, \ldots, \mathbf{y}_N$ that maximizes the penalized log-likelihood (9). Then,*

$$\sqrt{N}(\hat{\mathbf{x}}^{\text{PML}} - \check{\mathbf{x}}) \overset{a}{\sim} \mathcal{N}\left(0, (\mathbf{J}(\check{\mathbf{x}}) + \beta_0 \mathbf{M}(\check{\mathbf{x}}))^{-1} \mathbf{C}(\check{\mathbf{x}}) (\mathbf{J}(\check{\mathbf{x}}) + \beta_0 \mathbf{M}(\check{\mathbf{x}}))^{-1}\right),$$

where $\beta_0 = \lim_{N\to\infty} \beta_N/N$,

$$\check{\mathbf{x}} = \arg\max \{E\{\log p(\mathbf{y};\mathbf{x})\} - \beta_0 R(\mathbf{x})\};$$

$$\mathbf{C}(\check{\mathbf{x}}) = cov\left\{\frac{\partial \log p(\mathbf{y};\check{\mathbf{x}})}{\partial \mathbf{x}}\right\};$$

$$\mathbf{J}(\check{\mathbf{x}}) = -E\left\{\frac{\partial^2 \log p(\mathbf{y};\check{\mathbf{x}})}{\partial \mathbf{x}^2}\right\};$$

$$\mathbf{M}(\check{\mathbf{x}}) = \frac{\partial^2 R(\check{\mathbf{x}})}{\partial \mathbf{x}^2}.$$

5.2 The PML Estimator and the UCRLB

From Theorem 3, the asymptotic total variance of $\hat{\mathbf{x}}^{\text{PML}}$ is

$$\frac{1}{N}\text{Tr}\left((\mathbf{J}(\check{\mathbf{x}}) + \beta_0 \mathbf{M}(\check{\mathbf{x}}))^{-1} \mathbf{C}(\check{\mathbf{x}}) (\mathbf{J}(\check{\mathbf{x}}) + \beta_0 \mathbf{M}(\check{\mathbf{x}}))\right), \tag{10}$$

and the asymptotic bias gradient is $\mathbf{D}_{\text{PML}} = \partial\check{\mathbf{x}}/\partial\mathbf{x}_0 - \mathbf{I}$, where differentiating the expression for $\check{\mathbf{x}}$ we have that

$$\frac{\partial \check{\mathbf{x}}}{\partial \mathbf{x}_0} = (\mathbf{J}(\check{\mathbf{x}}) + \beta_0 \mathbf{M}(\check{\mathbf{x}}))^{-1} \frac{\partial}{\partial \mathbf{x}_0} E\left\{\frac{\partial \log p(\mathbf{y};\check{\mathbf{x}})}{\partial \mathbf{x}}\right\}. \tag{11}$$

With $\gamma = \mathbf{D}_{\text{PML}}^* \mathbf{D}_{\text{PML}}$, it follows from Theorem 1 that the total variance of any estimate of $\mathbf{x}_0$ with bias gradient $\mathbf{D}$ such that $\text{Tr}(\mathbf{D}^*\mathbf{D}) \le \text{Tr}(\mathbf{D}_{\text{PML}}^* \mathbf{D}_{\text{PML}})$ satisfies

$$C \ge \frac{\alpha^2}{N}\text{Tr}\left((\mathbf{I} + \alpha \mathbf{J}_1)^{-2} \mathbf{J}_1\right), \tag{12}$$

where $\alpha > 0$ is chosen such that

$$\text{Tr}\left((\mathbf{I} + \alpha \mathbf{J}_1)^{-2}\right) = \text{Tr}\left(\left(\frac{\partial \check{\mathbf{x}}}{\partial \mathbf{x}_0} - \mathbf{I}\right)^* \left(\frac{\partial \check{\mathbf{x}}}{\partial \mathbf{x}_0} - \mathbf{I}\right)\right), \tag{13}$$

and

$$\mathbf{J}_1 = E\left\{\left(\frac{\partial \log p(\mathbf{y}_1;\mathbf{x}_0)}{\partial \mathbf{x}}^* \frac{\partial \log p(\mathbf{y}_1;\mathbf{x}_0)}{\partial \mathbf{x}}\right)\right\}, \tag{14}$$

is the Fisher information from a single observation. Therefore, if we can choose $R(\mathbf{x})$ such that (10) is equal to the bound in (12), then the corresponding PML estimator achieves the UCRLB with average bias constraint. Similarly, from Theorem 2, the variance of any estimate of $\mathbf{x}_0$ with bias gradient $\mathbf{D}$ such that $\|\mathbf{D}\|^2 \le \|\mathbf{D}_{\text{PML}}\|^2$ satisfies

$$C \ge \frac{1}{N}\text{Tr}\left(\left(1 - \left\|\frac{\partial \check{\mathbf{x}}}{\partial \mathbf{x}_0} - \mathbf{I}\right\|\right)^2 \mathbf{J}_1^{-1}\right), \tag{15}$$

so that if we can choose $R(\mathbf{x})$ such that (10) is equal to the bound in (15), then the corresponding PML estimator achieves the UCRLB with worst-case bias constraint.

To develop intuition into the optimal choice of $R(\mathbf{x})$, we consider estimating a scalar x_0 from N iid measurements.

Theorem 4 *Let x_0 denote an unknown deterministic parameter, let $\mathbf{y}_1, \ldots, \mathbf{y}_N$ denote N iid vector measurements of x_0, and let $\hat{x}^{\text{PML}}$ denote the PML estimator of x_0 from the measurements $\mathbf{y}_1, \ldots, \mathbf{y}_N$ that maximizes the penalized log-likelihood with penalizing function $R(x)$. Then $\hat{x}^{\text{PML}}$ asymptotically achieves the UCRLB if and only if $R(x)$ is chosen such that*

$$\left(1 - \left|\frac{\partial \check{x}}{\partial x_0} - 1\right|\right)^2 \frac{1}{J_1} = \frac{C(\check{x})}{(J(\check{x}) + \beta_0 M(\check{x}))^2}.$$

In addition, if $\partial\check{x}/\partial x_0 \le 1$, then $\hat{x}^{\text{PML}}$ asymptotically achieves the UCRLB if and only if $R(x)$ is chosen such that

$$\frac{\partial \log p(\mathbf{y};\check{x})}{\partial x} - E\left\{\frac{\partial \log p(\mathbf{y};\check{x})}{\partial x}\right\} = c\frac{\partial \log p(\mathbf{y};x_0)}{\partial x}, \tag{16}$$

for some deterministic constant c.

In many cases, the condition (16) is satisfied for all $R(x)$, so that any $R(x)$ such that $\partial\check{x}/\partial x \le 1$ is asymptotically optimal. For example, suppose we are given measurements $\mathbf{y}_i = \mathbf{m} + \sigma_0 \mathbf{n}_i$, $1 \le i \le N$, where the mean, $\mathbf{m}$, is a known length-n vector, $\mathbf{n}_i$ are iid random vectors with $\mathbf{n}_1 \sim \mathcal{N}(0, \mathbf{I})$, and σ_0 is unknown. Then,

$$\frac{\partial \log p(\mathbf{y};\check{\sigma})}{\partial \sigma} = -\frac{n}{\check{\sigma}} + \frac{1}{\check{\sigma}^3}(\mathbf{y} - \mathbf{m})^*(\mathbf{y} - \mathbf{m}). \tag{17}$$

Since $E\{(\mathbf{y} - \mathbf{m})^*(\mathbf{y} - \mathbf{m})\} = n\sigma_0^2$, we have that

$$\frac{\partial \log p(\mathbf{y};\check{x})}{\partial x} - E\left\{\frac{\partial \log p(\mathbf{y};\check{x})}{\partial x}\right\} = \frac{\check{\sigma}^3}{\sigma_0^3}\frac{\partial \log p(\mathbf{y};x_0)}{\partial x}, \tag{18}$$

so that (16) is satisfied for all $R(x)$. The same conclusion holds when estimating $\mathbf{m}$, assuming σ_0 is known. Another, non-Gaussian example, is considered in the next section.

6. EXAMPLE

Consider the case in which we are given N scalar iid measurements $y_1,\ldots,y_N$ of an exponential random variable with unknown mean $1/x_0 > 0$, so that

$$p(y_i;x_0) = x_0 e^{-y_i x_0}, \quad 1 \le i \le N. \tag{19}$$

The PML estimate $\hat{x}^{\text{PML}}$ with penalizing function $R(x)$ is given by the value of x that maximizes

$$PL(x) = N\log x - x\sum_{i=1}^{N} y_i - \beta_N R(x), \tag{20}$$

for some $\beta_N > 0$ such that $\beta_N/N \to \beta_0$ as $N \to \infty$. We seek a penalizing function $R(x)$ that is optimal in the sense that the resulting estimator asymptotically achieves the UCRLB. (Note that in the scalar case, the average and worst-vase UCRLBs coincide.)

We can immediately verify that

$$\frac{\partial \log p(y;\check{x})}{\partial x} - E\left\{\frac{\partial \log p(y;\check{x})}{\partial x}\right\} = \frac{\partial \log p(y;x_0)}{\partial x_0}, \tag{21}$$

so that from Theorem 4 it follows that for any choice of $R(x)$ such that $\partial\check{x}/\partial x_0 \le 1$, the resulting PML estimator asymptotically achieves the UCRLB, where in this case

$$\frac{\partial \check{x}}{\partial x_0} = \frac{1/x_0^2}{1/\check{x}^2 + \beta_0 M(\check{x})}. \tag{22}$$

Note, however, that for finite values of N, the performance of the PML estimator will depend on the specific choice of penalizing function $R(x)$.

If $\partial R(\check{x})/\partial x, \partial^2 R(\check{x})/\partial x^2 \ge 0$, then

$$\frac{1}{\check{x}} = \frac{1}{x_0} + \beta_0 \frac{\partial R(\check{x})}{\partial x} \ge \frac{1}{x_0}, \tag{23}$$

so that $\partial\check{x}/\partial x_0 \le 1$, and the PML estimator is optimal. As an example, suppose that $R(x) = x$. Then the resulting PML estimator is

$$\hat{x}^{\text{PML}} = \frac{N}{\sum_{i=1}^{N} y_i + \beta_N}. \tag{24}$$

Since $\partial R(\check{x})/\partial x = 1 \ge 0$ and $\partial^2 R(\check{x})/\partial x^2 = 0$, it follows that the estimator of (24) asymptotically achieves the UCRLB.

As another example, suppose that $R(x) = \log x$. In this case, $\check{x} = (1-\beta_0)x_0$, so that from (22),

$$\frac{\partial \check{x}}{\partial x_0} = \frac{1/x_0^2}{(1-\beta_0)/\check{x}^2} = 1 - \beta_0 \le 1. \tag{25}$$

We therefore conclude that the resulting PML estimator, given by

$$\hat{x}^{\text{PML}} = \frac{N-\beta_N}{\sum_{i=1}^{N} y_i}, \tag{26}$$

asymptotically achieves the UCRLB.

In Fig. 1 we plot the UCRLB and the estimated variance of the PML estimators (24) and (26) as a function of the estimated squared bias gradient, for $N = 30$. The variance and the squared bias gradient of the estimators are computed using the method described in [3]. As we expect from our analysis, for increasing values of N the variance of both estimators approaches the UCRLB. Note, however, that in simulations it has been observed that for small values of N, the performance of the two estimators is different. In particular, simulations show that the estimator given by (24) results in a smaller variance than the estimator given by (26) for finite values of N.

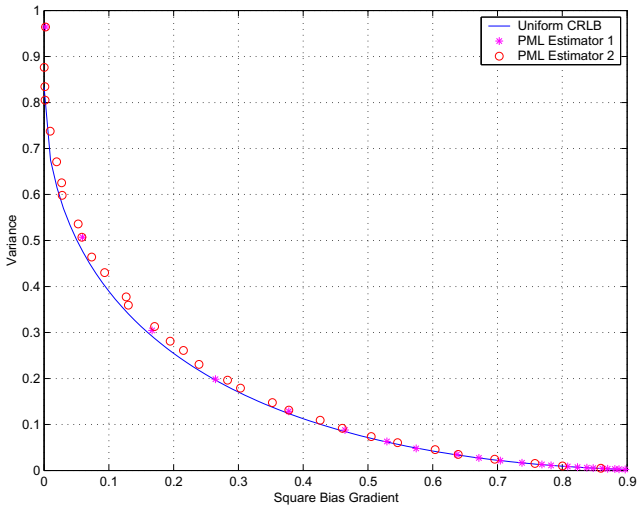

Figure 1: Performance of the PML estimators (24) (denoted "1") and (26) (denoted "2") with $N = 30$, in comparison with the UCRLB.

REFERENCES

[1] C. R. Rao, *Linear Statistical Inference and Its Applications*, New York, NY: John Wiley & Sons, Inc., second edition, 1973.

[2] H. L. Van Trees, *Detection, Estimation, and Modulation Theory*, John Willey and Sons, Inc., 1968.

[3] A. O. Hero, J. A. Fessler, and M. Usman, "Exploring estimator bias-variance tradeoffs using the uniform CR bound," *IEEE Trans. Signal Processing*, vol. 44, no. 8, pp. 2026–2041, Aug. 1996.

[4] Y. C. Eldar, "Minimum variance in biased estimation: Bounds and asymptotically optimal estimators," *IEEE Trans. Signal Processing*, to appear.

[5] L. Vandenberghe and S. Boyd, "Semidefinite programming," *SIAM Rev.*, vol. 38, no. 1, pp. 40–95, Mar. 1996.

[6] A. N. Tikhonov and V. Y. Arsenin, *Solution of Ill-Posed Problems*, Washington, DC: V.H. Winston, 1977.

[7] L. S. Mayer and T. A. Willke, "On biased estimation in linear models," *Technometrics*, vol. 15, pp. 497–508, Aug. 1973.

[8] J. A. Fessler and A. O. Hero, "Penalizes maximum-likelihood image reconstruction using space-alternating EM algorithms," *IEEE Trans. Image Processing*, vol. 4, pp. 1417–1425, Oct. 1995.

[9] J. A. Fessler, "Mean and variance of implicity defined biased estimators (such as penalized maximum likelihood): Applications to tomography," *IEEE Trans. Image Processing*, vol. 5, no. 3, pp. 493–506, Mar. 1996.